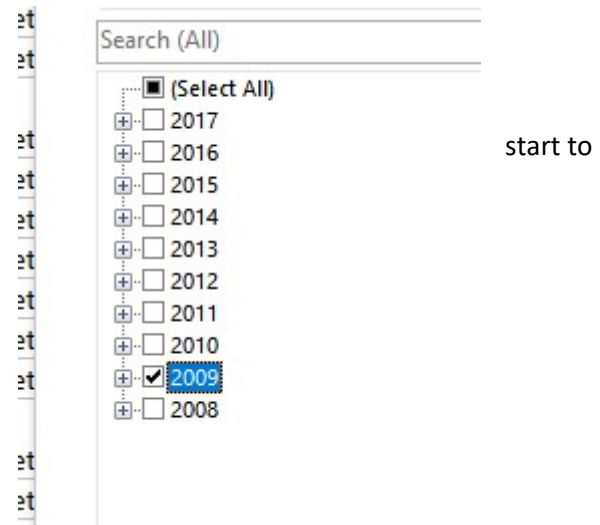


Statistics in Surveys

By Don Bremer

Under the date, select 2009

Copy the resulting data to a new sheet and we can run some analysis on it.



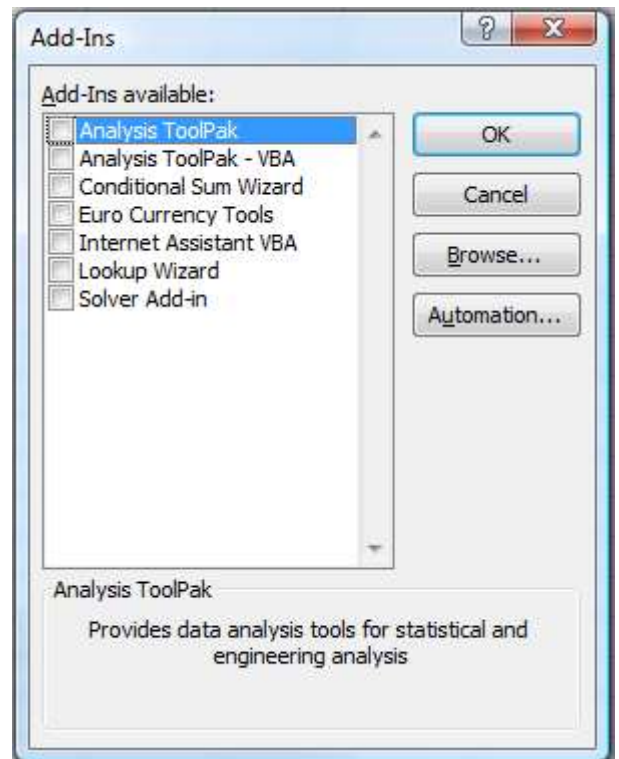
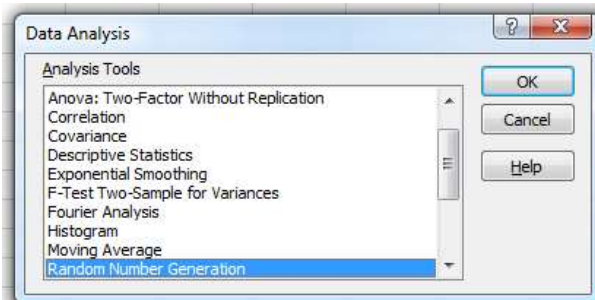
The data analysis toolpak

Even though it seems that the makers of Excel put the kitchen sink in and ready to go – not everything is visible. There is another statistical package that is an Add-In if you want to go even further.

File->Options->Add-Ins->Manage Excel Add-ins (Go...)

Select on Analysis ToolPak and click OK.

A new item appears under data called "Data Analysis". When the item is selected, it shows the different things you can do:



Let's select Descriptive Statistics. Let's see if I knew what I was talking about in 2009!

Select the data from \$I\$2:\$I\$81

Let's also select the Summary Statistics and then click ok.

Answers:

Column1	
Mean	4.8375
Standard Error	0.041505
Median	5
Mode	5
Standard Deviation	0.371236
Sample Variance	0.137816
Kurtosis	1.514995
Skewness	-1.86487
Range	1
Minimum	4
Maximum	5
Sum	387
Count	80

Descriptive Statistics

Input
 Input Range:
 Grouped By: ☒ Columns ☐ Rows
☐ Labels in First Row

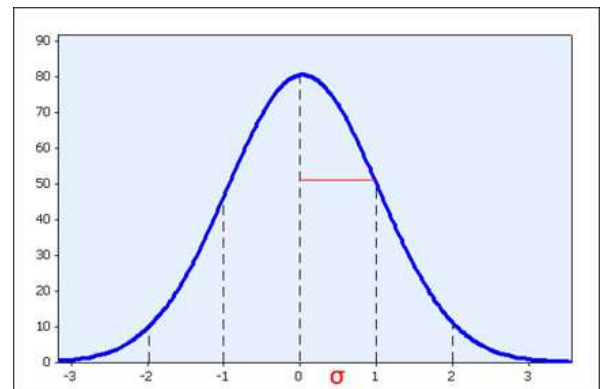
Output options
☐ Output Range:
☒ New Worksheet Ply:
☐ New Workbook
☒ Summary statistics
☐ Confidence Level for Mean: %
☐ Kth Largest:
☐ Kth Smallest:

Some definitions

- Mean – What we would call the average
- Median – The middle value if all the numbers were in order
- Mode – The number most often repeated
- Standard Error – A measure of the variability of the set. The deviation over number of items.
- Standard Deviation – used to find the amount of variation in a group of items
- Sample Variance - The average of the squared differences from the Mean.
- Kurtosis – A descriptor of the distribution. A normal distribution has a kurtosis of 0.
- Skewness - The measure of asymmetry in the distribution. A normal distribution has no skew (0).
- Range - The difference between the largest and smallest values.

This shows this was primarily 4s and 5s with a deviation of .37. This means the bulk of people are within .37 of a point of each other. This is a pretty good indication that it isn't just a random fluke.

Also, I see the kurtosis is pretty low (1.5). This is like the "tailedness" of the values. We will see what happens when this is high....



Let's compare this to Q10 (is this a convenient location?) – something I always think of as a throw away question –

This standard deviation is .72 – twice as much and nearer to 1 value away from typical. So, although good, it does show that my teaching is more consistent than how people thought of the location! (Whew!)

I also see that the Kurtosis is 19. That means the “tails” go out further in both directions. This means that the average answer on this varies quite a bit. As opposed the Range – which is just the difference between the minimum and maximum values.

By running this information on 2015 numbers, this is what we get:

<i>Column1</i>	
Mean	4.5875
Standard Error	0.080923761
Median	5
Mode	5
Standard Deviation	0.723804121
Sample Variance	0.523892405
Kurtosis	19.46622956
Skewness	3.497995401
Range	5
Minimum	0
Maximum	5
Sum	367
Count	80

Q6 – 2015		Q10- 2015	
<i>Column1</i>		<i>Column1</i>	
Mean	4.954545	Mean	4.560606
Standard Error	0.025836	Standard Error	0.112461
Median	5	Median	5
Mode	5	Mode	5
Standard Deviation	0.209895	Standard Deviation	0.913637
Sample Variance	0.044056	Sample Variance	0.834732
Kurtosis	18.51002	Kurtosis	9.300649
Skewness	-4.46652	Skewness	-2.74352
Range	1	Range	5
Minimum	4	Minimum	0
Maximum	5	Maximum	5
Sum	327	Sum	301
Count	66	Count	66

Am I getting better? Or, is the data just blurry? Or, are the students just getting smarter?

Types of Data

Qualitative (Attributes)

- Nominal
- Ordinal

Quantitative (Metrics)

- Numeric

Nominal Attributes

Data that be counted, but not ordered or aggregated (grouped into classes or clusters).

Examples:

- Products – Books, Movies, Music
- Gender – Male, Female
- State – Virginia, Nevada, California

What are some for your data?

Ordinal Attributes

Data that can be counted and ordered, but not aggregated

Examples:

- Date – 1/1/2014, 1/2/2014...
- Grades – A, B, C...
- Ranks – Like, Neutral, Dislike

What are some for your data?

Metrics

Quantitative data that can be counted, ordered, and aggregated.

Examples:

- Revenue, Cost, Profit
- Number of Customers
- Temperature
- Time

What are some for your data?

Ordinal Attributes and Metrics

Some data can be used as either attributes or metrics. Their classification is dependent on usage.

Examples:

- Age
- Scores

What are some for your data?

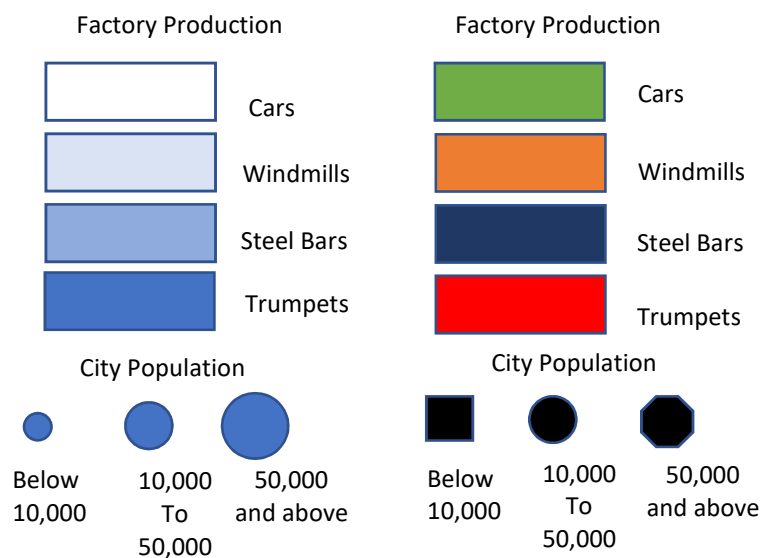
Visualizations

	Metric	Attribute (Ordinal)
Attribute (Nominal)	Bar Heatmap	Line (with Groups) Bar (with Groups)
Attribute (Ordinal)	Column Line	Scatter Grid
Metric	Scatter/ Bubble	

Appropriate Visual Enhancements

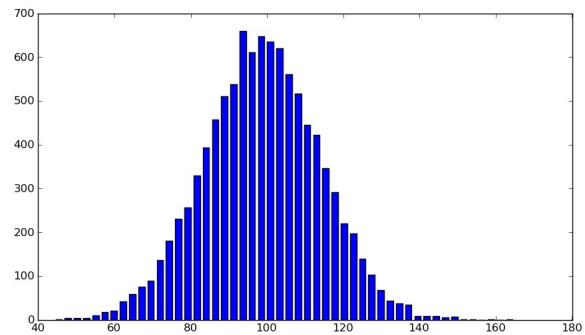
	Attribute (Nominal)	Attribute (Ordinal)	Metric
Color Hue	X	X	X
Color Saturation		X	X
Size		X	X

Use the right color scheme and icons for the right situation. Which icons or colors are better in the graphics below:



Histograms

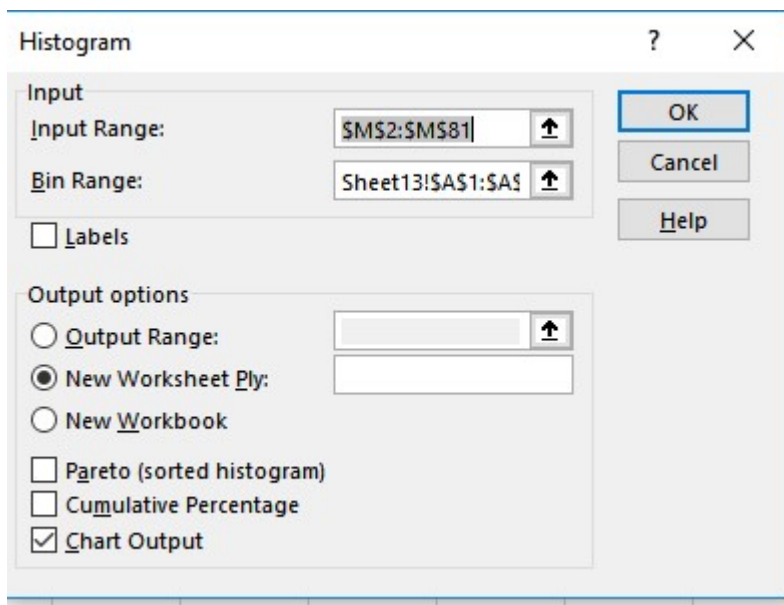
When quantitative data is what you have, a histogram would be used to show it. This is a kind of graph that also uses bars. Ranges of values are listed at the bottom and these are called 'classes.' Taller bars represent the classes with greater frequencies.



Create your own Histogram

Create a new sheet that has just the numbers 1-5. These will be the bins for the histogram.

Then, go to the 2009 data and start a histogram for Q10:



Histogram

Input

Input Range:

Bin Range:

☐ Labels

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

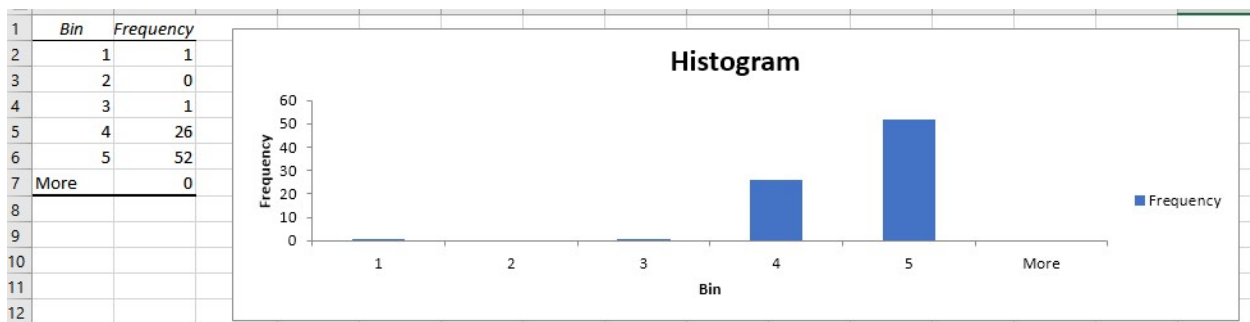
☐ Pareto (sorted histogram)

☐ Cumulative Percentage

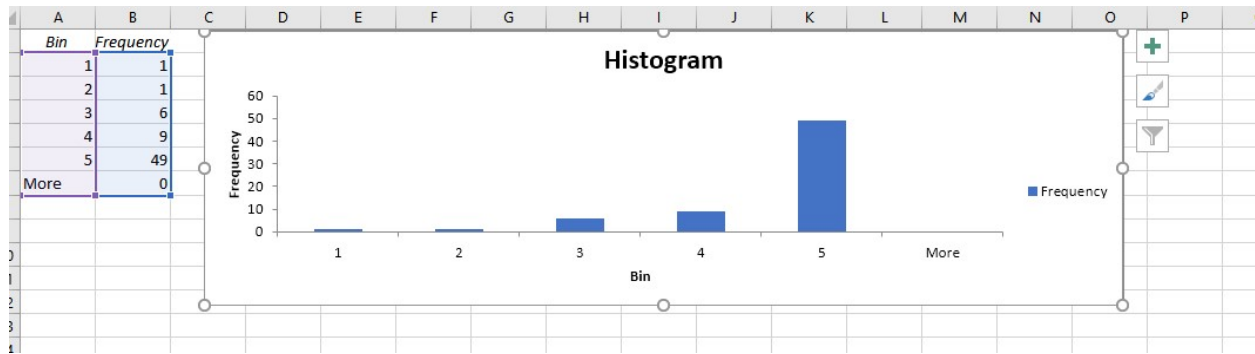
☒ Chart Output

OK Cancel Help

When this is executed, this is what we see:



Most are graphically 4 and 5 with a few small ones. Let's do it for 2015:



More of a spread. What does this tell us about the change in attitudes since 2009?

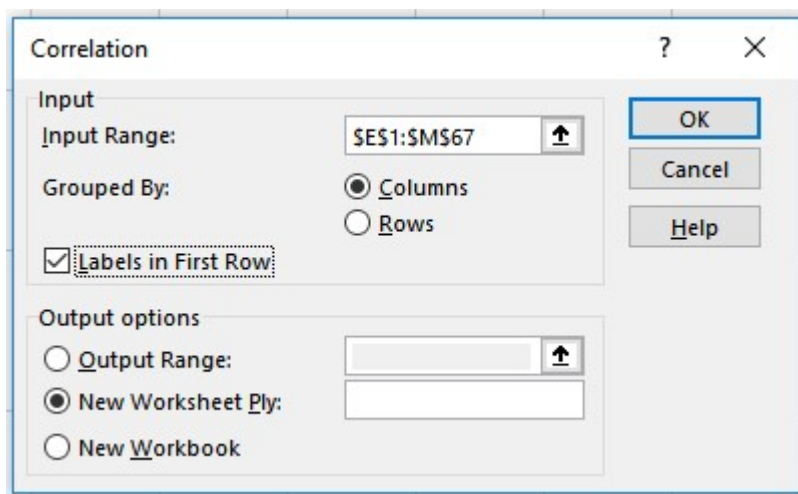
Statistics & Other Mind Blowing Items

Correlation

The correlation coefficient (a value between -1 and +1) tells you how strongly two variables are related to each other. We can use the CORREL function or the Analysis Toolpak add-in in Excel to find the correlation coefficient between two variables.

Note: A correlation coefficient of +1 indicates a perfect positive correlation. As variable X increases, variable Y increases. As variable X decreases, variable Y decreases.

Let's see if any of the questions correlate with one another when looking at the Excel class. Selecting all the data from the year 2015.



	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q2	1								
Q3	0.731399	1							
Q4	0.453875	0.646109	1						
Q5	0.633161	0.463254	0.525424	1					
Q6	0.440612	0.402374	0.259978	0.58757	1				
Q7	0.539937	0.601274	0.732004	0.623723	0.399255	1			
Q8	0.413424	0.580949	0.513259	0.326488	0.233759	0.572332	1		
Q9	0.457957	0.494266	0.421212	0.426746	0.46291	0.437682	0.444599	1	
Q10	0.138149	0.33532	0.338448	0.17843	0.134924	0.264218	0.236603	0.29147	1

This shows that there is a fairly high correlation (.73) between

- The information was presented effectively
- The seminar gave me good working knowledge of the subject matter

And nearly no correlation (.13) between

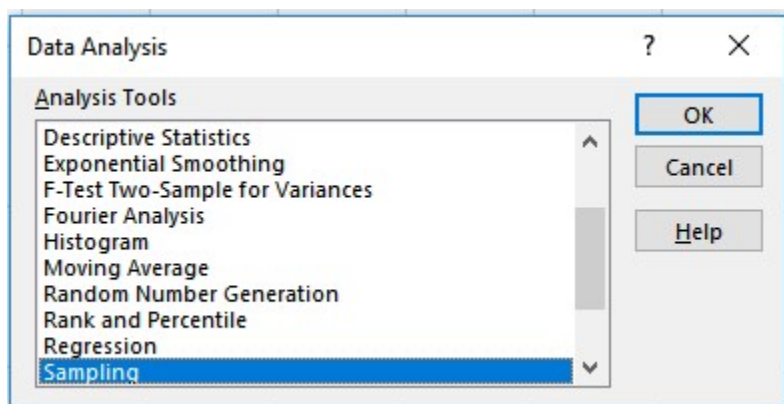
- The information was presented effectively
- The location was convenient.

I think this seems to track with what I would expect on these questions.

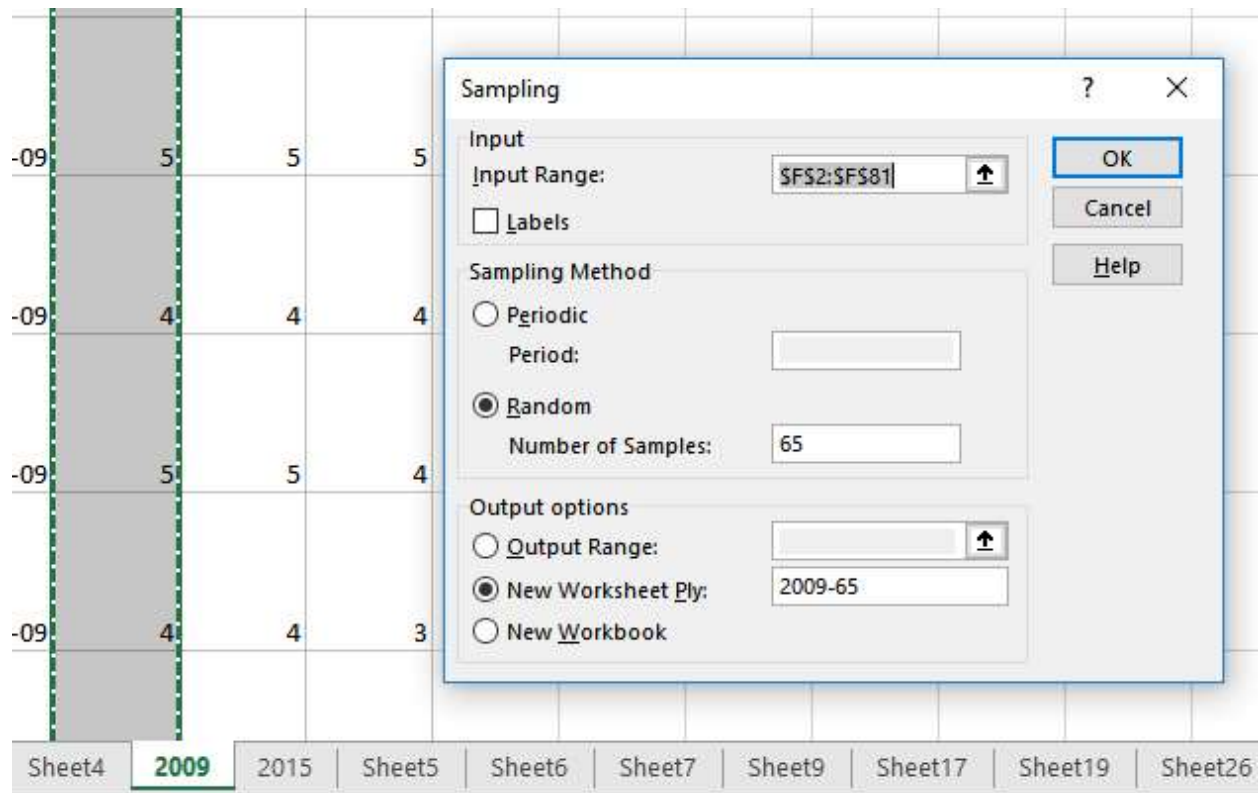
t-Test: Paired Two Sample for Means

This is the test to use to determine if two sets of data are *significantly* different from one another. So, now I will be able to tell if the changes that I have seen over the years are really statistically different from one another. This is usually used on the same set of subjects, but I'm turning this around and saying is **my** teaching getting better.

I'm going to grab 65 values from 2009. Why? Because I only have 65 values in 2015 and I need the same number of samples from both sets.



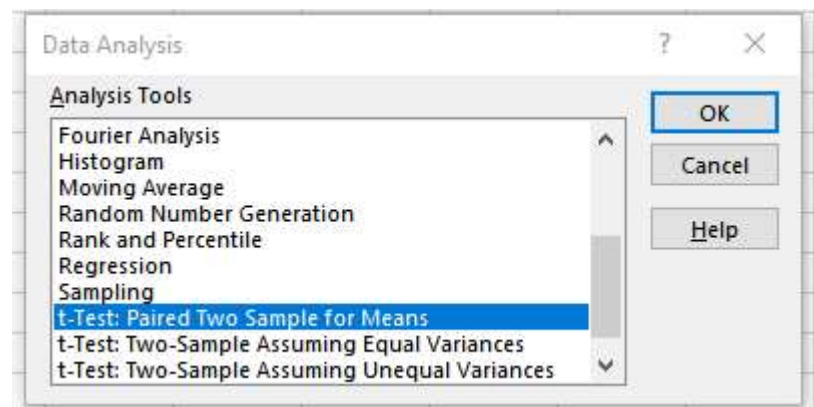
But, I don't want to have to pick just 65 samples. I would probably pick low samples so that it really looks like I got better! So, instead, I'm going to have the computer select 65 samples at random and put it on a new sheet – 2009-65.



Now, let see if I got better!

I'm going to make the assertion that: I have not gotten better in teaching over the last 6 years...

I'm going to select t-Test...



I'm going to use 2015 as Variable Range 1 and 2009 as Range 2.

t-Test: Paired Two Sample for Means

Input

Variable 1 Range: '2009'!\$F\$2:\$F\$67

Variable 2 Range: '2015'!\$E\$2:\$E\$67

Hypothesized Mean Difference: 0

☒ Labels

Alpha: 0.05

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

OK Cancel Help

Hypothesized Mean Difference – We are testing for 0 difference (whether I changed or not). You can also leave this blank if just testing for 0. If I wanted to say “Did I increase by .5”? – this is where I would put it.

Alpha – This is the change off of 95% confidence. We will keep it here for right now.

I’m going to put it in a new worksheet called t-Test 1 and click OK.

t-Test: Paired Two Sample for Means

	4	5
Mean	4.661538462	4.8
Variance	0.289903846	0.1625
Observations	65	65
Pearson Correlation	-0.172773685	
Hypothesized Mean Difference	0	
df	64	
t Stat	-1.537142693	
P(T<=t) one-tail	0.064593975	
t Critical one-tail	1.669013025	
P(T<=t) two-tail	0.12918795	
t Critical two-tail	1.997729654	

What do we see

In 2015 – My average was 4.8

In 2009 – My average was 4.66

But the bottom 4 stats are relevant for this test.

The Pearson Correlation tells if the two variables are correlated. It is close to zero – meaning there was really no correlation between the first test and the second test. This is ok – because it was actually different people!

The df is degrees of freedom. This is actually n-1 of the observation size.

tStat is what we are looking for and we are going to compare it to the t Critical statistic.

The t Critical stat is if we are saying that one number is below or one number is above. Since I picked no change (now you see why!) – I'm not going to use this... So, that's why we want to use the t Critical two tail..

A negative t-value simply indicates a reversal in the directionality of the effect, which has no bearing on the significance of the difference between groups. Analysis of a negative t-value requires examination of its absolute value in comparison to the value on a table of t-values and degrees of freedom (which quantifies the variability of the final estimated number). If the absolute value of the experimental t-value is smaller than the value found on the degrees of freedom chart, then the means of the two groups can be said to be significantly different.

Our t Stat (absolute) is actually smaller than the t Critical – so there is significance. How much?

I put both the values on one sheet and wanted to get the difference

	2015	2009	Diff
	5	5	=B3-A3
	5	4	=B4-A4
	5	4	=B5-A5
	5	4	=B6-A6
	5	5	=B7-A7
65	5	5	=B65-A65
66	5	5	=B66-A66
67	5	5	=B67-A67
68	5	5	=B68-A68
69			=AVERAGE(C3:C68)
70			

Name	Value	Description
Mean Difference	- .18182	The average difference is from 2009 to 2015 is - .18 (slightly worst in 2009)
Stand. Dev. Of Difference	.69433	Means that the standard dev of the differences is just under .7
Standard Error of Difference	0.086121	This is the Standard deviation divided by the sample size (where number matters)
T alpha half 95%	1.9977	From above
Lower Conf Level	-0.353862	Lower Conf. level of how bad I was
Upper Confidence Level	-0.009774	Higher Conf. level of how bad I was

	E	F	G
Mean Diff		-0.181818181818182	
Standard Dev		=STDEV(C3:C69)	
Standard Error of Diff		=F4/SQRT(65)	
T alpha half 95%		1.9977	
Lower confidence level		=F3-F5*F6	
Higher confidence level		=F3+F5*F6	

The statement is:

"I am 95% confident that my scores in 2009 were between -.0097 lower to -.353 lower than in 2015"

